Identifying regional dialects in online social media

Jacob Eisenstein

August 6, 2014

Electronic social media offers new opportunities for informal communication in written language, while at the same time, providing new datasets that allow researchers to document dialect variation from records of natural communication among millions of individuals. The unprecedented scale of this data enables the application of quantitative methods to automatically discover the lexical variables that distinguish the language of geographical areas such as cities. This can be paired with the segmentation of geographical space into dialect regions, within the context of a single joint statistical model — thus simultaneously identifying coherent dialect regions and the words that distinguish them. Finally, a diachronic analysis reveals rapid changes in the geographical distribution of these lexical features, suggesting that statistical analysis of social media may offer new insights on the diffusion of lexical change.

1 Dialect in social media

Social media comprises a wide range of different internet platforms, including collaborative writing projects such as Wikipedia, online communities such as Facebook and MySpace, forums such as Reddit and StackExchange, virtual game worlds, business and product reviews, and blogs and microblogs (Boyd and Ellison, 2007). These platforms offer a diverse array of ways to interact with friends and strangers — but in the overwhelming majority of cases, interaction is conducted in written language. As social media plays an increasingly ubiquitous part in daily life,[1] writing therefore acquires a new social role and a new level of importance. So it is unsurprising that a flowering of diversity in textual styles has been noticed by some observers (Walther and D'Addario, 2001; Crystal, 2006; Tagliamonte and Denis, 2008; Dresner and Herring, 2010; Collister, 2011; Schnoebelen, 2012) — as well as by some critics (Thurlow, 2006). Whether this newfound diversity is attributed to the lack of social regulation on social media writing (as compared with other written media), or to the demands of social, near-synchronous communication, it raises natural questions for dialectology: to what extent can geographical variation be observed in social media writing, how does it relate to geographical variation in spoken language, and what are the long-term prospects for this form of variation?

1.1 Twitter

Social media services vary in the degree to which large-scale data can be collected, and offer different sorts of geographic metadata. *Twitter* is a social media service that is particularly advantageous on both characteristics. Most messages are publicly-readable — this is the default setting — and can therefore be acquired through the streaming Application Programmer Inferface (API). In addition, many authors choose to include precise geographic coordinates as metadata with each message, although this is not enabled by default. In our data, roughly 1-2% of messages include geographical coordinates. Other methods for geolocating Twitter messages are explored by Dredze et al. (2013), but are not considered here.

Twitter is a *microblog* service, and messages ("tweets") are limited to 140 characters. The basic user interface is shown in Figure 1, and some examples of Twitter messages are shown here:

[1] Facebook reported 757 million daily active users in December 2013 (`http://investor.fb.com/releasedetail.cfm?ReleaseID=821954`, retrieved on April 5, 2014). Twitter reported 271 million monthly active users in August 2014 (`https://about.twitter.com/company`, retrieved on August 3, 2014).

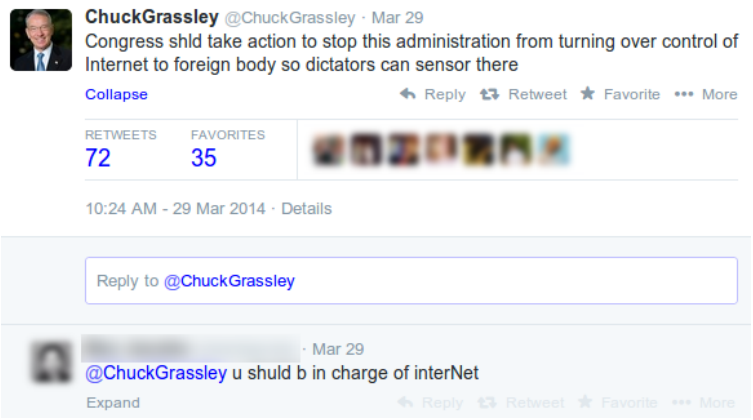


Figure 1: An example broadcast message followed by a conversational reply, which is addressed to the original author by beginning the message with his username.

bigdogcoffee Back to normal hours beginning tomorrow.....Monday-Friday 6am-10pm Sat/Sun 7:30am-10pm

crampell Casey B. Mulligan: Assessing the Housing Section - `http://nyti.ms/hcUKK9`

THE_REAL_SHAQ fill in da blank, my new years shaqalution is _____

These messages are chosen from public personae — a small business, a journalist, and a celebrity-athlete — but the majority of the content on Twitter is created by ordinary people. Users construct custom timelines by choosing to *follow* other users, whose messages then appear in the timeline. Unlike Facebook, these social network connections are directed, so that celebrities may have millions of followers (Kwak et al., 2010). For this reason, Twitter can be used as a broadcast medium. However, Twitter also enables dialogues in which messages can be "addressed" to another user by beginning the message with a username (Figure 1). This motivates another usage scenario, more akin to undirected social networks like Facebook, in which Twitter hosts public conversations (Huberman et al., 2008). More than 40% of messages in the dataset described below are addressed to another user.

Representativeness is a major concern with social media data. While social media increasingly reaches across barriers such as age, class, gender, and race, it cannot be said to offer a demographically balanced portrayal of the language community. Fortunately, the demographics of Twitter users in the United States have been surveyed repeatedly by the Pew Internet Research center (Duggan and Smith, 2013). Results from late 2013 found that 18% of Internet users visit Twitter, with nearly identical rates among men and women. Black Americans are significantly more likely to use Twitter than Whites, at a rate of 29% to 16% (for both Hispanic and non-Hispanic Whites); young people also use Twitter at a much higher rate (31% for ages 18-29, 19% for ages 30-49, and below 10% for older respondents). Differences across education level and income were not significant, but Twitter is used significantly less often in rural areas. An important caveat is that the *per-message* demographics may be substantially different from these figures, if the usage rate also varies with demographics. Consequently, it is important to remember that quantitative analysis of Twitter text (as with all social media) can describe only a particular demographic segment within any geographical area. This issue could be ameliorated through more fine-grained geographical analysis: for example, U.S. census blocks offer detailed demographic information, and their boundaries are drawn to emphasize demographic homogeneity (Eisenstein et al., 2011b). Another approach would be to try to infer the demographic characteristics of individual authors (Argamon et al., 2007; Chang et al., 2010; Rosenthal and McKeown, 2011), and then correct for these characteristics using post-stratification. Such refinements are beyond the scope of this chapter; their integration into social media dialectology must remain a topic for future work.

1.2 Related work

The computer science community has shown great interest in the problem of text-based geolocation: predicting where individuals are from, based on their writings (Cheng et al., 2010; Eisenstein et al., 2010; Wing and Baldridge, 2011; Hong et al., 2012). This task can be seen as the converse of the dialectologist's goal of summarizing the linguistic patterns that characterize residents of each geographical area. While predictive methods may yield insights for dialectology, the goals of accurate prediction and comprehensible modeling are not perfectly aligned, and therefore the most useful computational techniques may not necessarily be those which yield the most accurate predictions.

More broadly, several research efforts exploit social media datasets for purposes that touch on issues related to dialectology. Many researchers have attempted to model and predict the spread of online "memes" (Leskovec et al., 2009; Romero et al., 2011), and a related line of work investigates the survival of new words and expressions (Garley and Hockenmaier, 2012; Altmann et al., 2011). Other researchers have focused on the ways in which such temporal trends are shaped by groups, and on the emergence and evolution of linguistic conventions in online communities (Garley and Hockenmaier, 2012; Kooti et al., 2012; Nguyen and Rosé, 2011; Postmes et al., 2000; Danescu-Niculescu-Mizil et al., 2013b). A further consideration is the role of dyadic social relationships, which may shape linguistic behavior through phenomena such as accommodation (Danescu-Niculescu-Mizil et al., 2011), politeness (Danescu-Niculescu-Mizil et al., 2013a), power dynamics (Danescu-Niculescu-Mizil et al., 2012; Gilbert, 2012; Prabhakaran et al., 2012), and code-switching (Paolillo, 2011). Such research has mainly focused on the linguistic norms of online communities, such as forums and chatrooms, rather than on geographical regions in the physical world; it is geographically-anchored online language variation that constitutes the main focus on this chapter.

2 Dataset

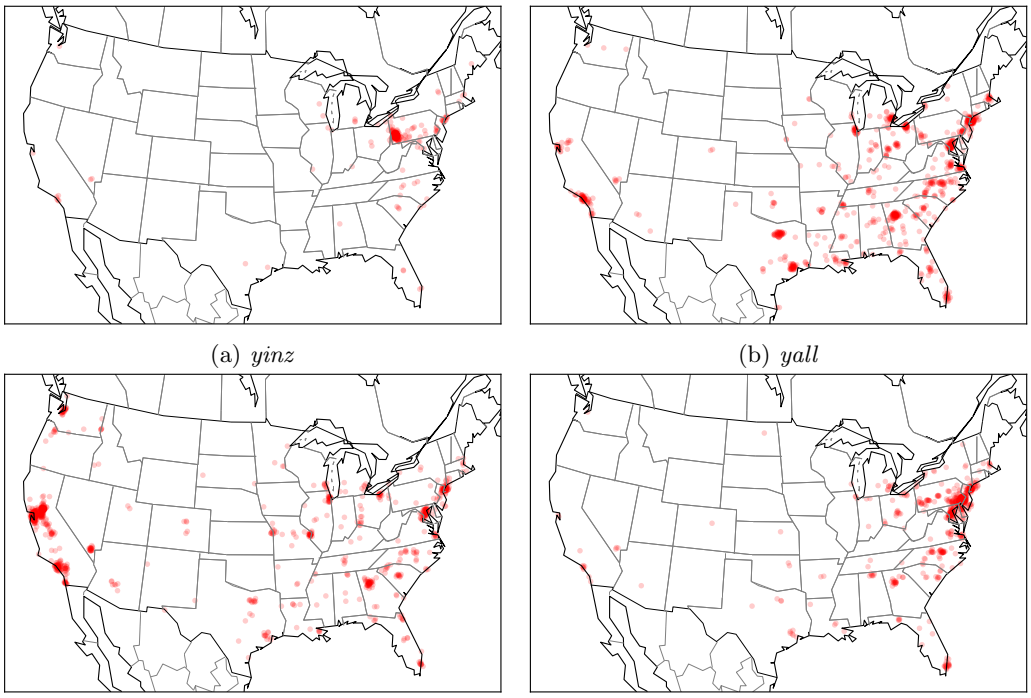
The empirical findings in this chapter are based on a dataset that was gathered by Brendan O'Connor from the public "Gardenhose" version of Twitter's streaming API (Application-Programmer Interface), and is first described in a technical report (Eisenstein et al., 2012). Within the initial set of messages, only those containing GPS metadata are considered here, so that analysis can be restricted to the United States. The streaming API ostensibly offers a 10% sample of public posts, although Morstatter et al. (2013) show that messages containing GPS metadata are sampled at a much higher rate. The dataset has been acquired by continuously receiving data from June 2009 to May 2012, and contains a total of 114 million geotagged messages from 2.77 million different user accounts.

Retweets are repetitions of previously-posted messages; they were eliminated using both Twitter metadata as well as the "RT" token (a common practice among Twitter users to indicate a retweet). Tweets containing URLs were eliminated in order to remove marketing-oriented messages, which are often automated. Accounts with more than 1000 followers or followees were removed for similar reasons. All text was downcased and tokenized using the publicly-available `Twokenize` program (Owoputi et al., 2013), but no other textual preprocessing was performed.

3 Known lexical variables

Before developing quantitative methods for discovering dialect variation in social media, I begin with a simpler question: do regional dialect words from spoken language persist in social media? This investigation will focus on four well-known spoken-language dialect terms.

- *Yinz* is a form of the second-person pronoun which is associated with the dialect of Southwestern Pennsylvania around the city of Pittsburgh (Johnstone et al., 2002). It is exceedingly rare in the Twitter dataset, appearing in only a few hundred messages, out of a total of one hundred million. The geographical distribution of these messages is indeed centered on the city of Pittsburgh, as shown in Figure 2a.
- *Yall* (also spelled *y'all*) is an alternative form of the second-person pronoun, often associated with the Southeastern United States, as well as African-American English (Green, 2002). It is relatively



(a) *yinz* (b) *yall*

(c) *hella* (d) *jawn*

Figure 2: Geolocations for messages containing four lexical variables known from previous work on spoken American English. In every case but *yinz*, 1000 randomly-selected examples are plotted; for *yinz*, only 535 examples are observed, so all are plotted.

frequent in the dataset, used at a rate of approximately one per 250 messages, making it more than 1000 times more common than *yinz*. Its geographical distribution, shown in Figure 2b, indicates that it is popular in the Southeast, but also in many other parts of the United States.

- *Hella* is an intensifier that is popularly associated with Northern California (Bucholtz et al., 2007); it is used in examples such as *i got hella nervous*. Unlike *yinz*, *hella* is fairly common in this dataset, appearing in nearly one out of every thousand messages. While the word does appear in Northern California at a higher-than-average rate, Figure 2c shows that it is used throughout the country.
- *Jawn* is a noun with Philadelphia origins (Alim, 2009) and diffuse semantics:

  (1) @name ok u have heard this jawn right

  (2) how long u been up in that jawn @name

  (3) i did wear that jawn but it was kinda warm this week

  Jawn appears at a rate of approximately one per ten thousand messages, with a geographical distributional reflecting its Philadelphia origins (Figure 2d). However, there is significant diffusion to nearby areas in New Jersey and New York City.

These maps show that existing lexical variables from spoken language do appear in social media text, and they are supported by analysis of other well-known variable pairs, such as *sub/hoagie* and *soda/pop*. However, variables that appear frequently have broad geographical distributions, which only loosely reflect their spoken-language geographical association (in the case of *hella*) or fail to reflect it altogether (*yall*). In the case of these two terms, this may be attributable to their popularity in African-American English, an ethnic dialect that is widely shared across the United States (Eisenstein et al., 2011b; Green, 2002). Conversely, terms with sharply-defined geographical signatures, such as *yinz* and *hoagie*, are exceedingly

rare. As noted above, Twitter users are on average younger and less likely to be White; terms such as *yinz* may be rarely used in this demographic. Only *jawn* maintains both a strong geographical signature and widespread popularity, and Alim (2009) traces this term's origins to African-American English.

4 Discovering lexical variables

The previous section shows that some lexical and phonological variables from spoken language have made the transition to online media, and that analysis of such media provides a large-scale but noisy picture of this variation. But the analysis thus far has been unsystematic, focusing on a few hand-chosen examples. A key advantage of big data is that it can speak for itself: we can apply data-driven statistical analysis to find words with strong geographical associations. This could reveal lexical variables from spoken language that were neglected by our intuition, or might identify variables that are new to electronic media.

Many researchers have examined the problem of automatically identifying words associated with types of documents or groups of authors. Monroe et al. (2008) provide a useful overview of relevant statistical techniques in this general area; Eisenstein et al. (2010) show how these methods can be applied to dialect analysis; in Chapter 24 of this volume, Grieve describes methods for measuring spatial autocorrelation in linguistic variables. Our approach here will be probabilistic, with the goal of characterizing how the word frequencies change with respect to geography. A simple first-pass approach would be to consider a set of metropolitan areas, and then compute the relative frequency of each term among all messages from within each area. Such frequencies can be computed directly from raw counts, but they are difficult to compare. If we consider the raw *difference* in frequencies, this will overemphasize very common words at the expense of rare ones; for example, an increase of 1% in the frequency of *the* or *and* would be far greater than the total frequency of *yinz* in Pittsburgh. If, on the other hand, we consider the *ratio* of frequencies, this will overemphasize rare words; in the limit, a word that appears just once in the entire dataset will have an infinite ratio of frequencies between two metropolitan areas.

One way to avoid these problems is to reparametrize the probability distribution over words, by applying the *logistic transformation*. This transformation takes the form of a ratio between two non-negative quantities, and the denominator ensures that the function sums to one over all words — thus satisfying the basic requirements of a probability distribution. Specifically, for each region r, we have:

$$P_r(w) = \frac{\exp(m_w + \beta_w^{(r)})}{\sum_i \exp(m_i + \beta_i^{(r)})}, \tag{1}$$

where $m_w = \log \hat{P}(w)$, the log of the empirical frequency of word w across the entire population (in all regions), and $\beta_w^{(r)}$ is the *deviation* from this empirical log frequency in region r. The numerator exponentiates this sum, ensuring non-negativity, and the denominator sums over all words, ensuring that $\sum_w P_r(w) = 1$. Assuming $\boldsymbol{\beta}^{(r)}$ is centered at zero, then a large positive value $\beta_w^{(r)}$ means that word w is substantially more frequent in region r than it is elsewhere; a large negative value means it is less frequent. Sorting by $\beta_w^{(r)}$ proves a very effective way of identifying meaningful words for region r, striking a good balance between uniqueness and frequency. *Comprehensibility* is therefore the first of two advantages for this formulation. Similar observations were made by Monroe et al. (2008), in the context of analyzing speech differences between political parties in the United States.

Since $\boldsymbol{m}$ is computed directly from the empirical frequencies, the *estimation* problem is to compute each parameter vector $\boldsymbol{\beta}^{(r)}$. We will use a regularized maximum-likelihood criterion, meaning that we choose $\boldsymbol{\beta}^{(r)}$ to maximize the log-likelihood of the observed text, $\mathcal{L} = \sum_n \log P(w_n; \boldsymbol{\beta}^{(r)}, \boldsymbol{m})$, subject to a *regularizer* that penalizes $\boldsymbol{\beta}^{(r)}$ for its distance from zero. Regularization (also called shrinkage) reduces the sensitivity of the parameter estimates to rare terms in the data, by introducing a bias towards uniformity (Murphy, 2012). Because of this bias, strong evidence is required before we have $\beta_w^{(r)} \neq 0$; therefore, *robustness* is the second advantage of this formulation. A typical choice for the regularizer is to penalize a norm of the vector $\boldsymbol{\beta}^{(r)}$. More details, including an algorithm to to estimate $\boldsymbol{\beta}^{(r)}$, can be found in a prior publication (Eisenstein et al., 2011a). Source code is also freely available.[2]

[2] https://github.com/jacobeisenstein/SAGE

This approach can be applied to the dataset described in Section 2, with each region corresponding to a *metropolitan statistical area* (MSA). MSAs are defined by the United States government, and include the regional area around a single urban core. Using this approach, the top words for some of the largest MSAs in the United States are:

- **New York**: flatbush, baii, brib, bx, staten, *mta*, odee, soho, deadass, werd
- **Los Angeles**: pasadena, venice, anaheim, *dodger*, disneyland, angeles, compton, *ucla*, *dodgers*, melrose
- **Chicago**: #chicago, lbvs, chicago, *blackhawks*, *#bears*, *#bulls*, mfs, *cubs*, burbs, bogus
- **Philadelphia**: jawn, ard, *#phillies*, *sixers*, *phils*, *wawa*, *philadelphia*, delaware, *philly*, *phillies*

The plurality of these terms are place names (underlined) and geographically-specific entities (italicized), such as sports teams (*dodgers*, *sixers*), businesses (*wawa*, a grocery store), and local government agencies (*mta*, which is responsible for mass transportation in New York City). However, there are several other types of words, which are of greater interest for dialect.

- **Dialect words from speech**. The term *jawn* was already discussed as a feature of spoken Philadelphia English, and the terms *burbs* (suburbs) and *bogus* (fake) may also be recognized in spoken language. The term *deadass* — typically meaning "very", as in *deadass serious* — may be less familiar, and might have passed unnoticed without the application of automated techniques.
- **Alternative spellings**. The spelling *werd* substitutes for the term *word* — but only in the senses identified by Cutler (1999), as in *oh, werd?* (*oh really?*), or as affirmation, as in *werd, me too*. Note that the spelling *word* is also used in these same contexts, but the spelling *werd* is almost never used in the standard sense.

  More remotely, *ard* is an alternative spelling for *alright*, as in:

  (4) @name ard let me kno

  (5) lol (*laugh out loud*) u'll be ard

  Similarly, *brib* is an alternative spelling for *crib*, which in turn signifies *home*.

  (6) bbq (*barbecue*) at my fams (*family's*) brib

  (7) in da brib, just took a shower

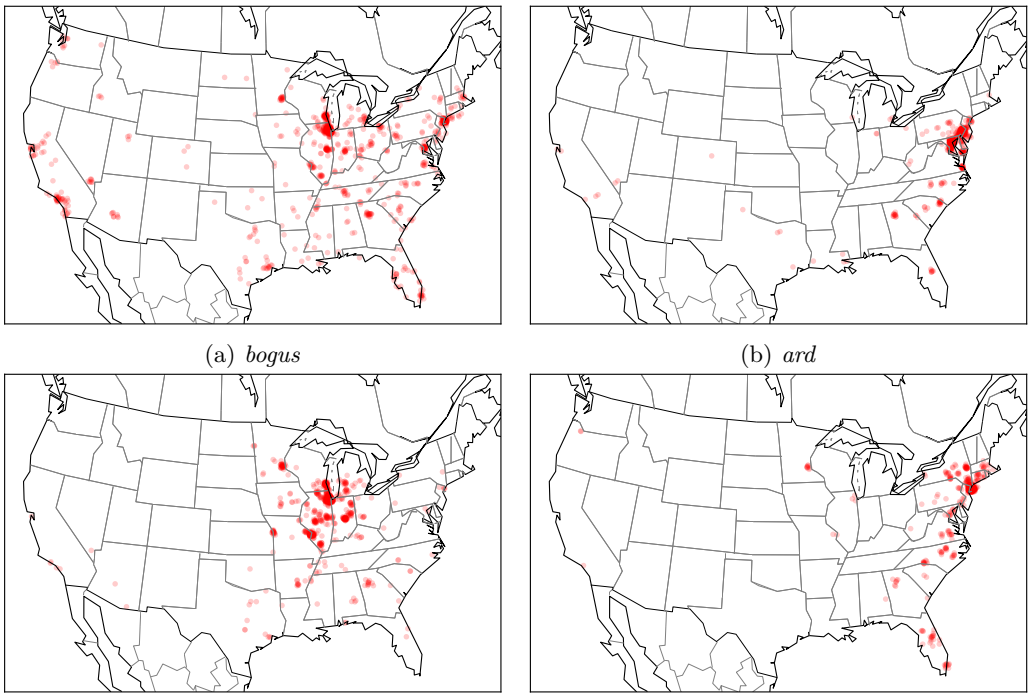
  Nationally, *brib* appears at a rate of once per 22,000 messages, which is roughly 5% as often as *crib*. But in the New York City area, *brib* appears at a rate of once per 3,000 messages.

  A final example is *baii*, which is an alternative to *boy*, in the sense of a friend or partner. As shown in the second example below, it may also function as a pragmatic marker: similar to *man* in Cheshire's (2013) study of urban English in the U.K., it can be used without referring to any specific individual.

  (8) look at my baii @name congrats again wish i was there 2 see u walk baii

  (9) i'm outta here baii

- **Abbreviations**. The abbreviation *lol* (*laugh out loud*) is well-known in the discourse about social media text, but several lesser-known abbreviations have strong regional affiliations. These include *lbvs* (*laughing but very serious*) and *mfs* (*motherfuckers*, as in *these mfs are crazy*). My prior work with collaborators at Carnegie Mellon University (Eisenstein et al., 2010) identified several other phrasal abbreviations with non-uniform geographical distributions, including *af* (an intensifier signifying *as fuck*), *ctfu* (*cracking the fuck up*), and *lls* (*laughing like shit*).
- **Combinations** A few words appear to combine aspects of multiple types. The word *odee* is a phonetic spelling of the abbreviation *od*, which stands for *overdose*, but it is now used as an intensifier with considerable syntactic flexibility.



(a) *bogus* (b) *ard*

(c) *lbvs* (d) *odee*

Figure 3: Four examples of lexical variables discovered from social media analysis

(10) she said she odee miss me

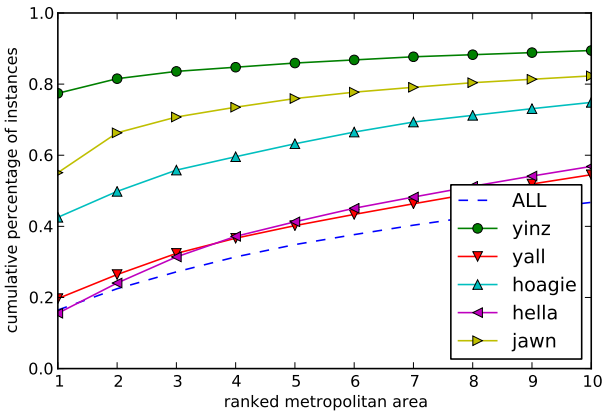
(11) its rainin odee :(

(12) for once i'm odee sleepy

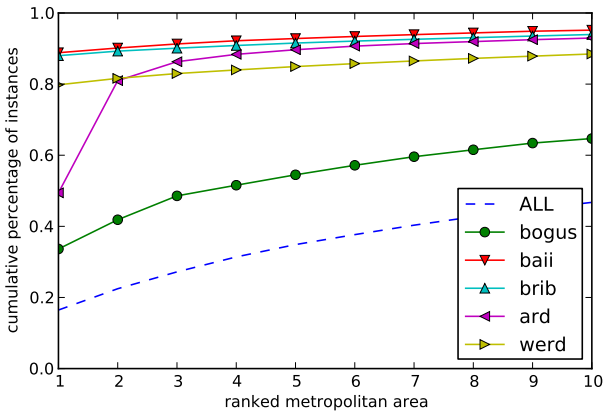
The geographical distributions of four of these terms are shown in Figure 3. Another measure of the regional specificity of these words can be seen in the cumulative fraction of word counts accounted for by their home cities (Figure 4). For example, for the word *yinz*, 77% of the counts come from Pittsburgh; for *ard*, 81% of the counts come from its top two cities of Philadelphia and Baltimore. Each subfigure also includes the overall cumulative proportion of word counts, indicating that 16% of all counts come from the largest metropolitan area, New York, and that this fraction increases slowly to nearly 50% when the ten largest metropolitan areas are considered. This line is what one would expect to see for words with no significant geographical association, and indeed, Figure 4a shows *yall* and *hella* track it closely. Figures 4b and 4c show that many of the strongest geographical orientations belong to automatically-discovered social media terms — particularly to terms associated with New York, such as *baii*, *brib*, and *odee*. Figure 4d shows that the geographical associations for these social media terms are stronger than the corresponding associations for many place and entity names.

5 Discovering dialect regions

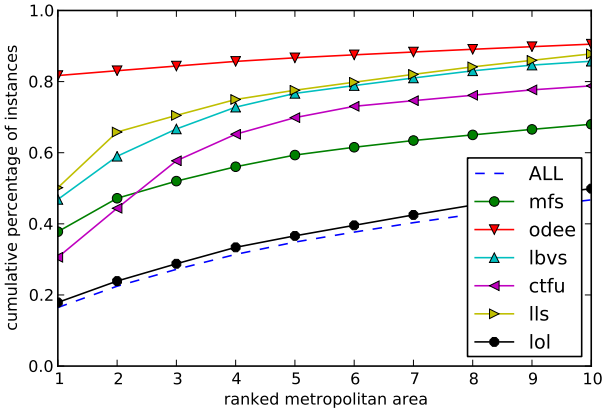
Figure 4 shows that many words are strongly associated with individual metropolitan areas — indicated by a high initial value and then a plateau in the graph. But some words, such as *ard*, *lls*, and *ctfu* have a steeper slope for the first few points — indicating that they are strong in two or more metropolitan areas. This suggests that interesting dialect regions may span multiple metropolitan areas. Moreover, such a view could reveal different linguistic patterns, since the words we are currently analyzing were chosen on the basis of their association with individual metropolitan areas. By grouping metropolitan areas, we are more likely to notice variables that belong to smaller cities, which individually might not have large enough word counts to survive the statistical shrinkage applied to individual metropolitan areas.



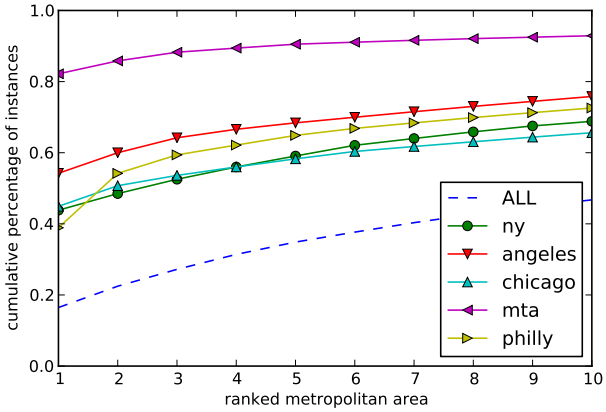
(a) Previously-known spoken-language dialect words



(b) Automatically discovered social media dialect terms



(c) Automatically discovered social media abbreviations



(d) Toponyms and local entities

Figure 4: Cumulative proportion of counts across top metropolitan areas. The plot for a word will be flat to the extent that its usage is dominated by a single city, and will track the "all" line to the extent that its usage is independent of geography.

Other corpus-based approaches to dialectometry have applied clustering to identify dialect regions from matrices of *distances* between more fine-grained geographical areas (Heeringa and Nerbonne, 2001; Szmrecsanyi, 2011). Such distance matrices are typically computed from catalogs of predefined dialect features; for example, Szmrecsanyi (2011) uses a set of 57 morphosyntactic features, including the frequency of alternative tense markings and systems of negation. More details on these approaches can be found in Chapters 18 and 23 of this volume. In social media, the relevant dialect features may not be known in advance; part of the goal is to discover them from raw text. While it is still possible to compute distance functions — for example, by applying a divergence metric between the word frequency distributions in two metropolitan areas (Wing and Baldridge, 2011) — such distance metrics can be difficult to apply when data is sparse and high-dimensional, because they are sensitive to the amount and nature of smoothing applied to the word counts.

Recall that the overall goal is to partition the map into a set of regions, which are both spatially compact and linguistically consistent. We can frame this objective probabilistically: we want to choose a set of regions and regional word distributions so that the observed data — the text and its geolocations — have a high likelihood under some reasonable probability model. To do this, we treat the regions, and the assignment of authors to regions, as *latent variables*. We then define a probabilistic model that unifies the latent variables with the observed text and geolocations (Eisenstein et al., 2010). Bayesian inference in this model combines the likelihood of the observed data with our prior beliefs about the latent variables (e.g., the typical size of a geographical region), using the machinery of Bayes' law (Murphy, 2012; Nerbonne, 2007).

How can we define a probabilistic model that unifies the observations and latent variables? Our approach will be "generative," in that we propose a fictional stochastic process by which the latent variables and observations are generated. This generative process is not intended as a psycholinguistic model of writing; it is simply a way to arrange the variables so that the quantities of interest can be recovered through statistical inference. There are three key quantities: the assignment of authors to dialect regions, the spatial extent of the dialect regions, and the word distributions associated with each region.[3] The relationships between these quantities can be formalized by assigning them variable names, and proposing statistical dependencies between the variables, as shown in Algorithm 1.

for *author* $a \in \{1 \dots A\}$ **do**
  Randomly sample a *region* $z_a \sim P(z; \theta)$, where θ defines a prior distribution over regions. For example, θ will assign high likelihood to the region containing New York, and low likelihood to more sparsely-populated regions.
  for *each tweet* $t \in \{1 \dots T_a\}$ **do**
    Randomly sample a *geolocation* $\boldsymbol{x}_{a,t} \sim P(\boldsymbol{x}_{a,t}; \phi_{z_a})$, where $\boldsymbol{x}_{a,t}$ is a latitude-longitude pair and ϕ_{z_a} specifies a distribution over such pairs. The form of the distribution over latitude and longitude pairs is Gaussian in prior work (Eisenstein et al., 2010). In general the Gaussian distribution is inappropriate for latitude and longitude on a sphere, but it is a suitable approximation for a small section of the globe, such as the continental United States.
    for *each word token* $n \in \{1 \dots N_{t,a}\}$ **do**
      Randomly sample a *word* $w_{a,t,n} \sim P(w; \boldsymbol{m}, \boldsymbol{\beta}_{z_a})$, using the probability distribution defined in Equation 1.
    end
  end
end

Algorithm 1: Stochastic generative model for geolocated text

The goal of statistical inference is to obtain estimates of the quantities of interest, under which the observed words and geolocations attain high probability. Specifically, we must infer the assignment of authors to regions (z), the prior distribution over regions (θ), the geographical extent of each region (ϕ), and the word distributions for each region ($\boldsymbol{\beta}$). By summarizing the probabilistic dependencies between the latent and observed variables, Algorithm 1 provides the basic specification for an inference procedure. Toolkits for *probabilistic programming* are capable of automatically transforming such a specification directly into executable code for statistical inference, without requiring any manual derivations or programming (Lunn

[3]The models described by Eisenstein et al. (2010) and Eisenstein et al. (2011a) are more ambitious, attempting to distinguish latent *topics*, which capture variation that is independent of geography. This is beyond the scope of this chapter.

et al., 2000; Goodman et al., 2008; Stan Development Team, 2014).[4] However, such toolkits were not used in this work, because this model admits relatively straightforward inference through the application of variational expectation maximization (Wainwright and Jordan, 2008). The details of this procedure are beyond the scope of this chapter, and are described in prior publications (Eisenstein et al., 2010, 2011a). Variational expectation maximization is similar to soft K-means clustering, alternating between two steps: (1) making soft assignments of authors to regions (clusters), and (2) updating the linguistic and geographic centroids of the clusters. Eventually, this procedure converges at a local optimum.

6 Change over time and other next steps

The analysis thus far has been entirely synchronic, but social media data can also shed light on how online language changes over time. Many of the terms mentioned in previous sections, such as *ctfu, ard, baii*, are nearly unknown in English writing prior to the recent era of computer-mediated communication. Yet dialect lexicons have expanded so rapidly that these terms are now in use among several thousands of individuals, and in some cases, across wide areas of the United States.

Figure 5 shows some examples. The term *af* (*as fuck*), is used mainly in Southern California and Atlanta in 2010, but attains widespread popularity by 2012. The term *ion* (meaning *i don't*; it is very rarely used in the chemical sense in this dataset) appears in a few scattered Southern cities in 2010, but spreads widely throughout the South by 2012. The emoticon -__- (indicating ambivalence or annoyance) was popular in several of the largest coastal urban areas in 2010 — with remarkably limited popularity in the interior metropolises of Chicago, Dallas, and Houston — but reached widespread urban popularity by 2011, and nearly universal usage by 2012. These examples all offer support for various versions of the gravity model of linguistic change (Trudgill, 1974), in which new features spread first between the most populated cities, with a limited role for geographical diffusion. However, other examples are more difficult to explain in terms of population alone: for example, *ctfu* (*cracking the fuck up*) spreads from its origin in Cleveland to adjacent parts of Pennsylvania, and from there to coastal cities along the mid-Atlantic; it does not spread westward to the large, nearby metropolises of Detroit and Chicago until much later. The abbreviation *lbvs* (*laughing but very serious*) is used almost exclusively in Chicago in 2010, and becomes popular in several other Midwestern cities by 2012 — though not yet the nearby cities of Detroit and Cleveland. The phonetic spelling *ard* is highly popular in Baltimore and Philadelphia, but does not spread to the neighboring city of Washington DC — a distance of 90 kilometers. Several of the terms most strongly associated with New York City (*odee*, *werd*, *deadass*) also fail to attain much popularity outside their city of origin.

The search for explanations beyond population size and geographical proximity leads inevitably to considerations of race, class, and cultural differences, and these topics have long been seen as central to sociolinguistic research on language variation and change (Gordon, 2000; Labov, 2011). This strain of sociolinguistics has focused primarily on sound changes such as the Northern Cities Shift (Labov, 1994), but the slow pace of sound change and the limited number of measurable linguistic variables pose challenges for the distillation of a quantitative "grand unified theory" that accounts for geography, population size, and demographics. In this sense, social media data offers unique advantages: change is rapid enough to be measurable in real time, and the methods described in this chapter can be used to identify hundreds of linguistic features whose frequency correlates with both time and place. While such correlational analysis has already shown that many lexical features have strong demographic profiles (Eisenstein et al., 2011b), the role of demographics in the diffusion of these terms has yet to be definitively captured, though this is a topic of ongoing investigation. As Labov (1994) shows, the tightly interconnected nature of the phonological system means that a single sound change can cause a series of other changes, which may take several generations to play out. This property would seem to be lacking from lexical change — although the lexicon features its own patterns of interdependence (Pierrehumbert, 2010) — and if so, this would limit the extent to which generalizations can be drawn between these two systems. But from a social perspective, it may be precisely this relative simplicity that makes the lexicon the ideal laboratory in which to disentangle the complex social phenomena that regulate language variation and change.

Social media data may have still more to tell us about dialect. A clear priority is to export these methods to dialects outside the United States, particularly to places where variation is better described by

[4] A repository of software packages is found at `http://probabilistic-programming.org`

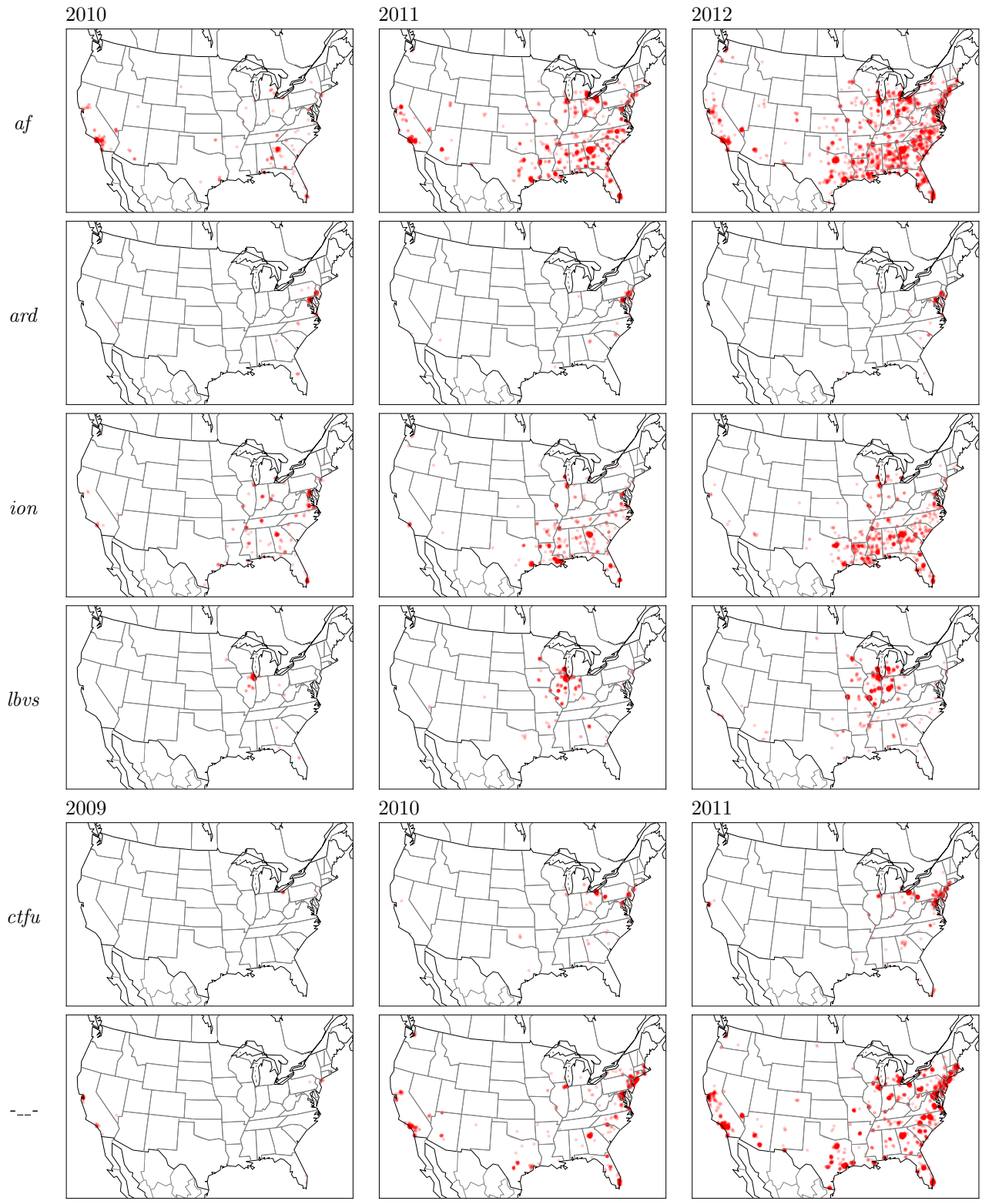


Figure 5: Geolocations for messages containing the words *af* (*as fuck*), *ard* (*alright*), *ion* (*i don't*), *lbvs* (*laughing but very serious*), *ctfu* (*cracking the fuck up*), and the emoticon -__- (ambivalence or annoyance).

continua rather than discrete regions (Heeringa and Nerbonne, 2001), which may require new computational methods. Another topic of interest is the relationship between written and spoken dialect: to what extent is phonological regional variation transcribed into written language in social media? Preliminary work suggests that several phonological variables are often transcribed in Twitter writing (Eisenstein, 2013), but these phenomena (such as "g-dropping") mainly correlate with register varation, rather than with geographical dialects. Still another key question is how the use of geographical variables is affected by other properties of the author: for example, younger people may write more informally, but older authors may be more likely to use traditional variables such as *yinz*. Finally, sociolinguistics has been increasingly concerned with the role of language variation in fine-grained conversational situations (Jaffe, 2012) — a phenomenon that is very difficult to measure at scale without social media data. Preliminary work suggests that in conversational dialogues, authors modulate the frequency with which they use local variables depending on properties of their interlocutor (Pavalanathan and Eisenstein, 2014), lending support to theories such as accommodation (Giles et al., 1991) and audience design (Bell, 1984). It is hoped that further studies in this direction will shed light on how dialect is perceived, and how it is deployed to create and reflect social relationships.

Acknowledgments

Thanks to Brendan O'Connor for providing the data on which this chapter is based, and for many insightful conversations over a fun and productive long-term collaboration on this research. Thanks are also due to John Nerbonne, Shawn Ling Ramirez, and Tyler Schnoebelen for providing helpful editorial suggestions on this chapter. This work benefitted from collaborations and discussions with David Bamman, Scott F. Kiesling, Brendan O'Connor, Umashanthi Pavalanathan, Noah A. Smith, Tyler Schnoebelen, and Eric P. Xing, and was supported by a grant from the National Science Foundation.

Biographical note

Jacob Eisenstein is Assistant Professor in the School of Interactive Computing at the Georgia Institute of Technology, where he leads the Computational Linguistics Laboratory. He received a doctorate in Computer Science from the Massachusetts Institute of Technology in 2008.

References

Alim, H. S. (2009). Hip hop nation language. In Duranti, A., editor, *Linguistic Anthropology: A Reader*, pages 272–289. Wiley-Blackwell, Malden, MA.

Altmann, E. G., Pierrehumbert, J. B., and Motter, A. E. (2011). Niche as a determinant of word fate in online groups. *PloS one*, 6(5):e19009.

Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9).

Bell, A. (1984). Language style as audience design. *Language in Society*, 13(2):145–204.

Boyd, D. and Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230.

Bucholtz, M., Bermudez, N., Fung, V., Edwards, L., and Vargas, R. (2007). Hella nor cal or totally so cal? the perceptual dialectology of california. *Journal of English Linguistics*, 35(4):325–352.

Chang, J., Rosenn, I., Backstrom, L., and Marlow, C. (2010). epluribus: Ethnicity on social networks. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, volume 10, pages 18–25.

Cheng, Z., Caverlee, J., and Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the Conference on Information and Knowledge Management (CIKM)*, pages 759–768.

Cheshire, J. (2013). Grammaticalisation in social context: The emergence of a new English pronoun. *Journal of Sociolinguistics*, 17(5):608–633.

Collister, L. B. (2011). *-repair in online discourse. *Journal of Pragmatics*, 43(3):918–921.

Crystal, D. (2006). *Language and the Internet*. Cambridge University Press, second edition.

Cutler, C. A. (1999). Yorkville crossing: White teens, hip hop and African American English. *Journal of Sociolinguistics*, 3(4):428–442.

Danescu-Niculescu-Mizil, C., Gamon, M., and Dumais, S. (2011). Mark my words! linguistic style accommodation in social media. In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 745–754.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., and Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 699–708.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., and Potts, C. (2013a). A computational approach to politeness with application to social factors. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 250–259.

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., and Potts, C. (2013b). No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 307–318.

Dredze, M., Paul, M. J., Bergsma, S., and Tran, H. (2013). Carmen: A Twitter geolocation system with applications to public health. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using Artificial Intelligence*, pages 20–24.

Dresner, E. and Herring, S. (2010). Functions of the nonverbal in CMC: Emoticons and illocutionary force. *Communication Theory*, 20(3):249–268.

Duggan, M. and Smith, A. (2013). Social media update 2013. Technical report, Pew Research Center.

Eisenstein, J. (2013). Phonological factors in social media writing. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 11–19.

Eisenstein, J., Ahmed, A., and Xing, E. P. (2011a). Sparse additive generative models of text. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1041–1048.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 1277–1287.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2012). Mapping the geographical diffusion of new words. Technical Report 1210.5268, ArXiV.

Eisenstein, J., Smith, N. A., and Xing, E. P. (2011b). Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1365–1374.

Garley, M. and Hockenmaier, J. (2012). Beefmoves: dissemination, diversity, and dynamics of English borrowings in a German hip hop forum. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 135–139.

Gilbert, E. (2012). Phrases that signal workplace hierarchy. In *Proceedings of Computer-Supported Cooperative Work (CSCW)*, pages 1037–1046.

Giles, H., Coupland, J., and Coupland, N. (1991). *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press.

Goodman, N. D., Mansinghka, V. K., Roy, D. M., Bonawitz, K., and Tenenbaum, J. B. (2008). Church: A language for generative models. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, pages 220–229.

Gordon, M. J. (2000). Phonological correlates of ethnic identity: Evidence of divergence? *American Speech*, 75(2):115–136.

Green, L. J. (2002). *African American English: A Linguistic Introduction*. Cambridge University Press.

Heeringa, W. and Nerbonne, J. (2001). Dialect areas and dialect continua. *Language Variation and Change*, 13(3):375–400.

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., and Tsioutsiouliklis, K. (2012). Discovering geographical topics in the twitter stream. In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 769–778.

Huberman, B., Romero, D. M., and Wu, F. (2008). Social networks that matter: Twitter under the microscope. *First Monday*, 14(1).

Jaffe, A., editor (2012). *Stance: Sociolinguistic Perspectives*. Oxford Studies in Sociolinguistics. Oxford University Press.

Johnstone, B., Bhasin, N., and Wittkofski, D. (2002). "Dahntahn" Pittsburgh: Monophthongal /aw/ and Representations of Localness in Southwestern Pennsylvania. *American Speech*, 77(2):148–176.

Kooti, F., Yang, H., Cha, M., Gummadi, P. K., and Mason, W. A. (2012). The emergence of conventions in online social networks. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 194–201.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 591–600.

Labov, W. (1994). *Principles of Linguistic Change*, volume 1: Internal Factors. Blackwell Publishers.

Labov, W. (2011). *Principles of Linguistic Change*, volume 3: Cognitive and Cultural Factors. Wiley-Blackwell.

Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of Knowledge Discovery and Data Mining (KDD)*, pages 497–506.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337.

Monroe, B. L., Colaresi, M. P., and Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's Streaming API with Twitter's Firehose. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 400–408.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Nerbonne, J. (2007). The exact analysis of text. In Mosteller, F. and Wallace, D., editors, *Inference and Disputed Authorship: The Federalist Papers*. CSLI: Stanford, third edition.

Nguyen, D. and Rosé, C. P. (2011). Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 76–85.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 380–390.

Paolillo, J. C. (2011). Conversational codeswitching on usenet and internet relay chat. *Language@Internet*, 8(3).

Pavalanathan, U. and Eisenstein, J. (2014). Linguistic style-shifting in online social media. In *review*.

Pierrehumbert, J. B. (2010). The dynamic lexicon. In Cohn, A., Huffman, M., and Fougeron, C., editors, *Handbook of Laboratory Phonology*, pages 173–183. Oxford University Press.

Postmes, T., Spears, R., and Lea, M. (2000). The formation of group norms in computer-mediated communication. *Human communication research*, 26(3):341–371.

Prabhakaran, V., Rambow, O., and Diab, M. (2012). Predicting overt display of power in written dialogs. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 518–522.

Romero, D. M., Meeder, B., and Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 695–704.

Rosenthal, S. and McKeown, K. (2011). Age prediction in blogs: A study of style, content, and online behavior in pre- and Post-Social media generations. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 763–772.

Schnoebelen, T. (2012). Do you smile with your nose? Stylistic variation in Twitter emoticons. *University of Pennsylvania Working Papers in Linguistics*, 18(2):14.

Stan Development Team (2014). Stan: A c++ library for probability and sampling, version 2.2.

Szmrecsanyi, B. (2011). Corpus-based dialectometry: a methodological sketch. *Corpora*, 6(1):45–76.

Tagliamonte, S. A. and Denis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. *American Speech*, 83(1):3–34.

Thurlow, C. (2006). From statistical panic to moral panic: The metadiscursive construction and popular exaggeration of new media language in the print media. *Journal of Computer-Mediated Communication*, pages 667–701.

Trudgill, P. (1974). Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society*, 3(2):215–246.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305.

Walther, J. and D'Addario, K. (2001). The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review*, 19(3):324–347.

Wing, B. and Baldridge, J. (2011). Simple supervised document geolocation with geodesic grids. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 955–964.